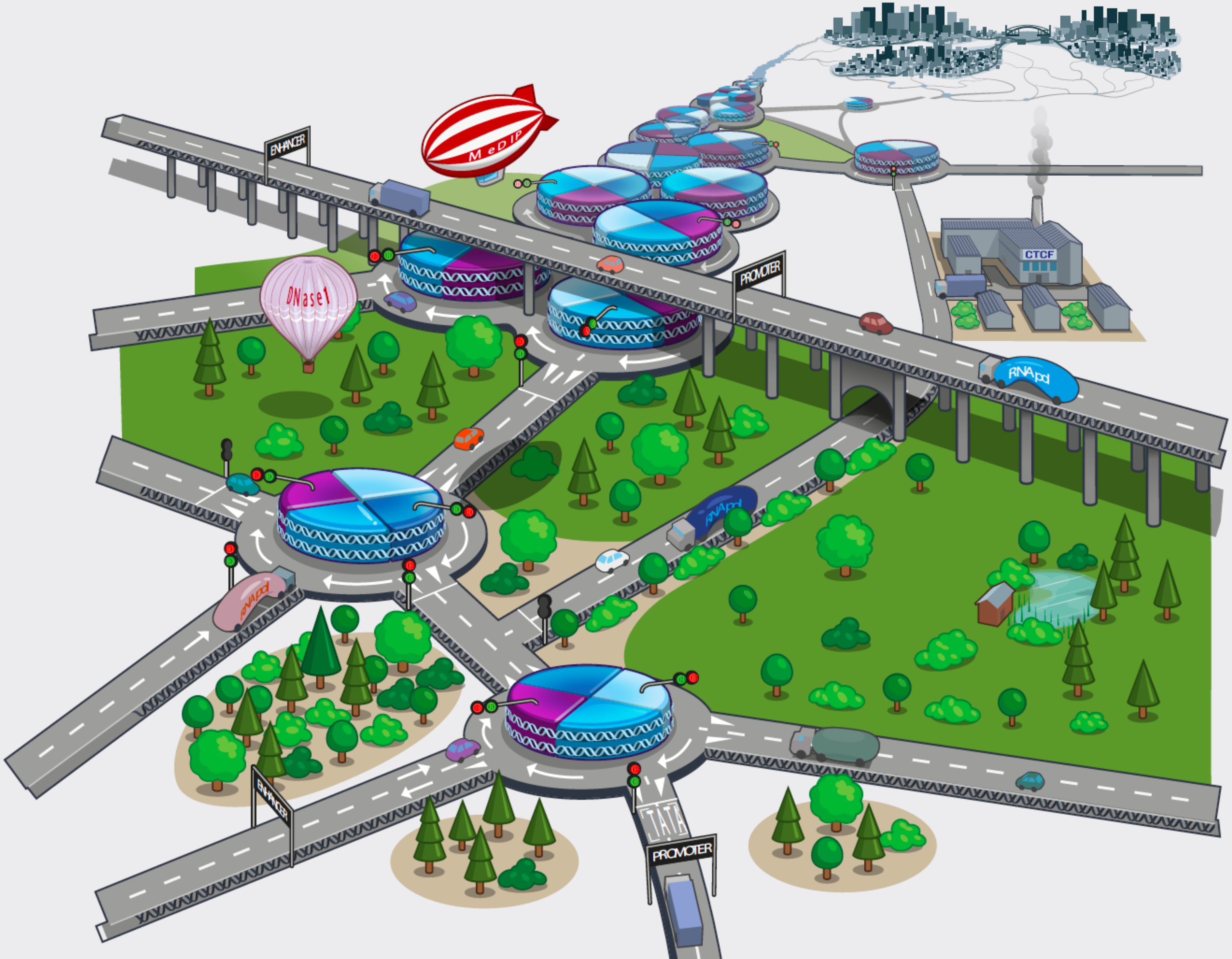
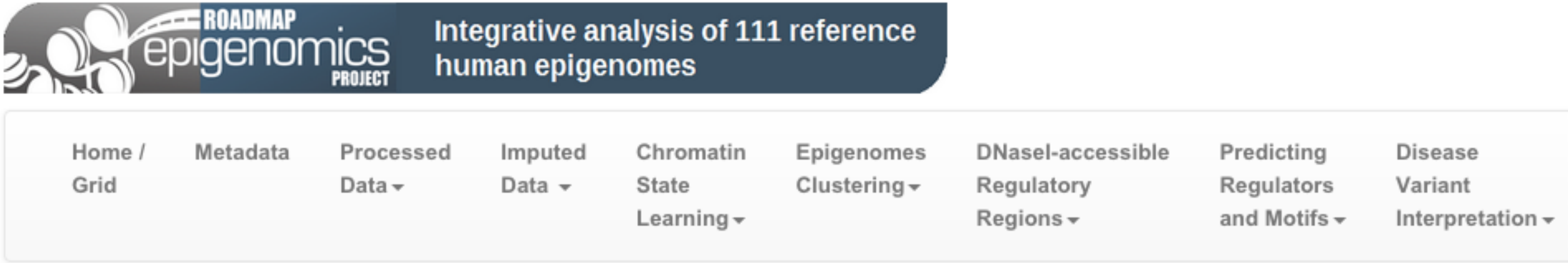


Overview of the supplementary website of the Roadmap Epigenomics project

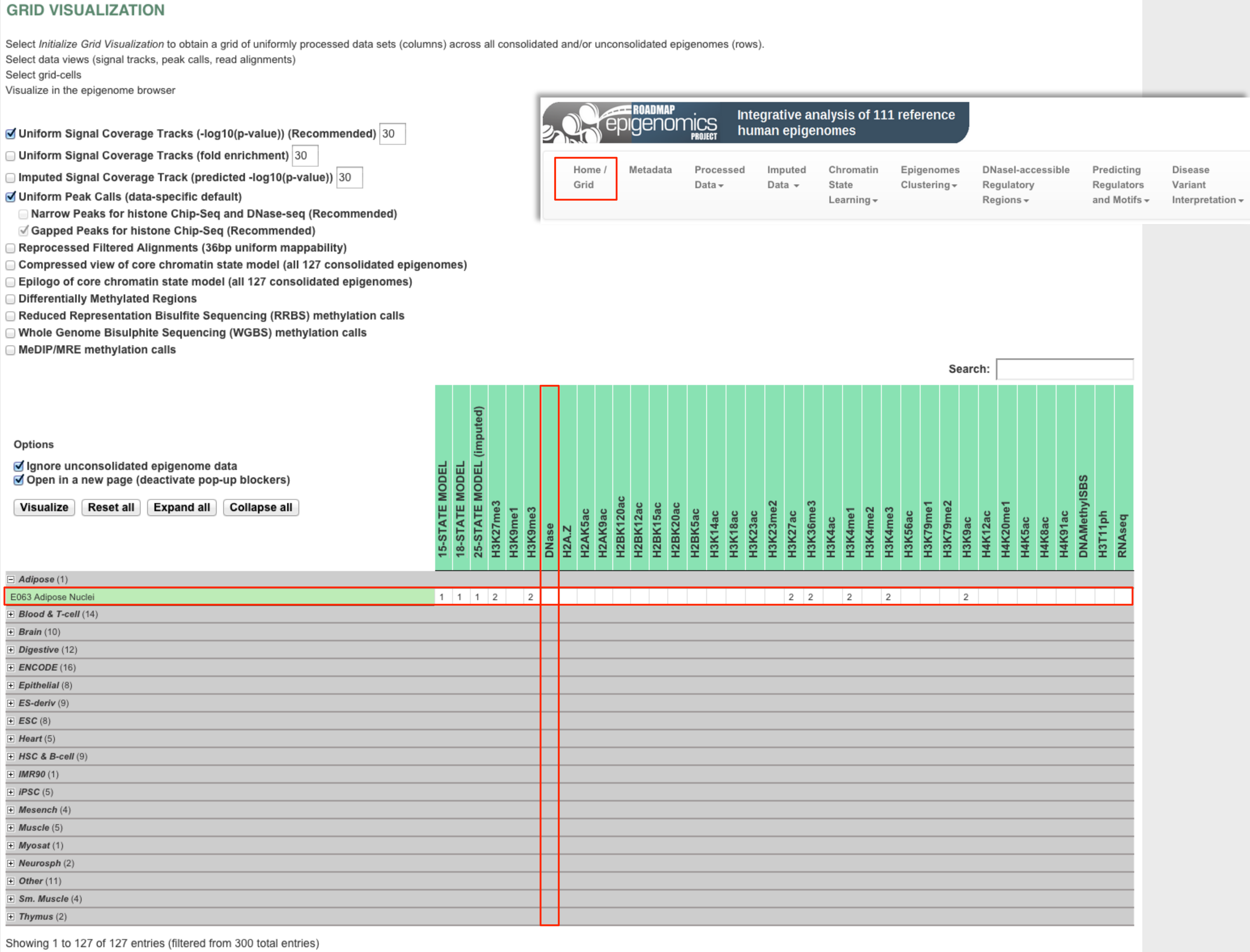
Wouter Meuleman
http://compbio.mit.edu/roadmap



Integrative analysis of 111 reference human epigenomes

- Website with supplementary information for the Roadmap Epigenomics Project
- Data (raw and uniformly processed)
- Annotations
- Methods

http://compbio.mit.edu/roadmap



GRID VISUALIZATION

Select *Initialize Grid Visualization* to obtain a grid of uniformly processed data sets (columns) across all consolidated and/or unconsolidated epigenomes (rows).
Select data views (signal tracks, peak calls, read alignments)
Select grid-cells
Visualize in the epigenome browser

- ☑ Uniform Signal Coverage Tracks (-log10(p-value)) (Recommended) 30
- ☐ Uniform Signal Coverage Tracks (fold enrichment) 30
- ☐ Imputed Signal Coverage Track (predicted -log10(p-value)) 30
- ☑ Uniform Peak Calls (data-specific default)
  - ☐ Narrow Peaks for histone Chip-Seq and DNase-seq (Recommended)
  - ☑ Gapped Peaks for histone Chip-Seq (Recommended)
- ☐ Reprocessed Filtered Alignments (36bp uniform mappability)
- ☐ Compressed view of core chromatin state model (all 127 consolidated epigenomes)
- ☐ Epilogo of core chromatin state model (all 127 consolidated epigenomes)
- ☐ Differentially Methylated Regions
- ☐ Reduced Representation Bisulfite Sequencing (RRBS) methylation calls
- ☐ Whole Genome Bisulphite Sequencing (WGBS) methylation calls
- ☐ MeDIP/MRE methylation calls

Search:

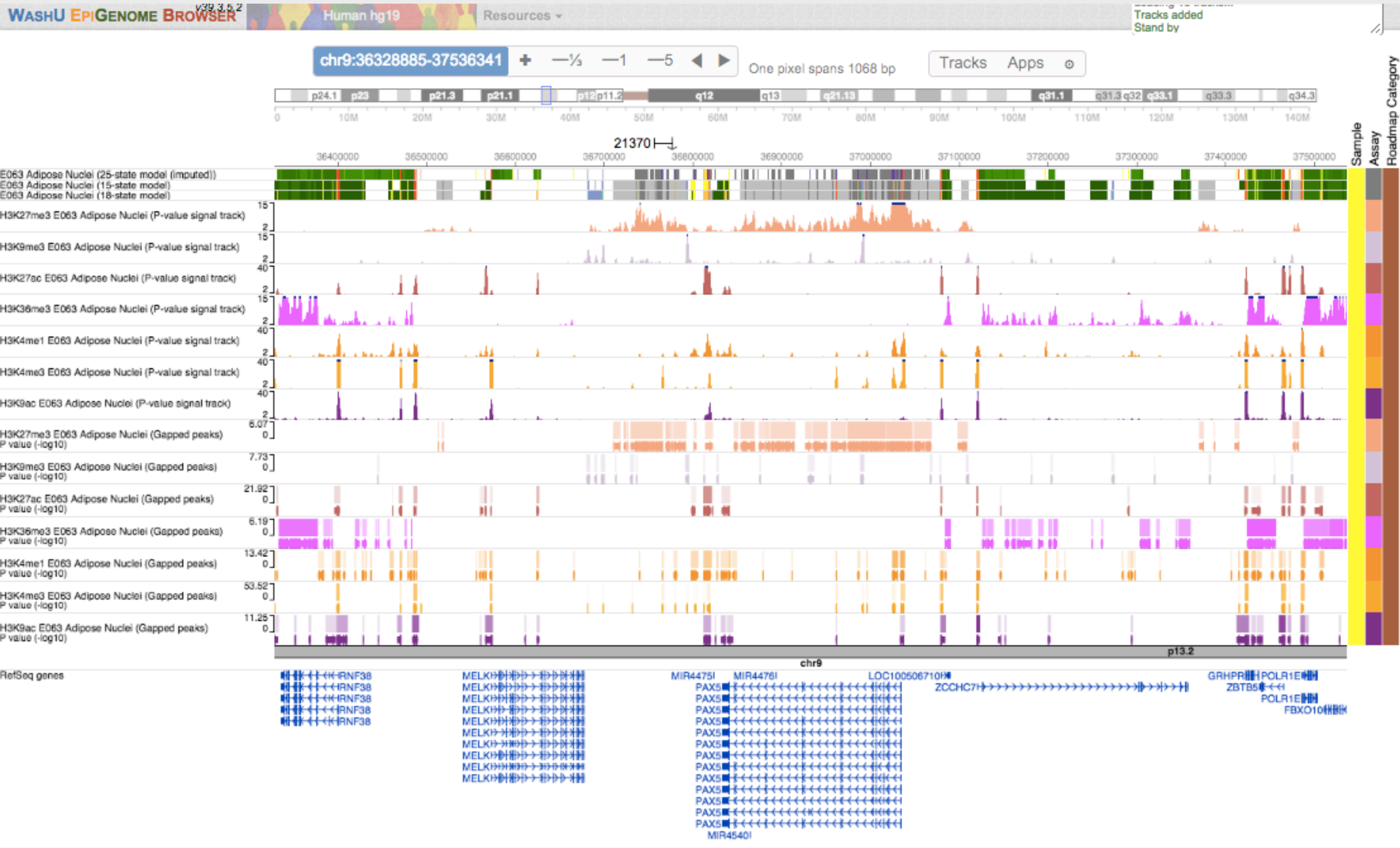
Options

- ☑ Ignore unconsolidated epigenome data
- ☑ Open in a new page (deactivate pop-up blockers)

	15-STATE MODEL	18-STATE MODEL	25-STATE MODEL (imputed)	H3K27me3	H3K9me1	H3K9me3	DNase	H2A.Z	H2AK5ac	H2AK9ac	H2BK120ac	H2BK12ac	H2BK15ac	H2BK20ac	H2BK5ac	H3K14ac	H3K18ac	H3K23ac	H3K23me2	H3K27ac	H3K36me3	H3K4ac	H3K4me1	H3K4me2	H3K4me3	H3K56ac	H3K79me1	H3K79me2	H3K9ac	H4K12ac	H4K20me1	H4K5ac	H4K8ac	H4K91ac	DNAMethylSBS	H3T11ph	RNAseq
⊟ ***Adipose*** (1)																																					
E063 Adipose Nuclei	1	1	1	2		2														2	2		2		2				2								
⊞ ***Blood & T-cell*** (14)																																					
⊞ ***Brain*** (10)																																					
⊞ ***Digestive*** (12)																																					
⊞ ***ENCODE*** (16)																																					
⊞ ***Epithelial*** (8)																																					
⊞ ***ES-deriv*** (9)																																					
⊞ ***ESC*** (8)																																					
⊞ ***Heart*** (5)																																					
⊞ ***HSC & B-cell*** (9)																																					
⊞ ***IMR90*** (1)																																					
⊞ ***iPSC*** (5)																																					
⊞ ***Mesench*** (4)																																					
⊞ ***Muscle*** (5)																																					
⊞ ***Myosat*** (1)																																					
⊞ ***Neurosph*** (2)																																					
⊞ ***Other*** (11)																																					
⊞ ***Sm. Muscle*** (4)																																					
⊞ ***Thymus*** (2)																																					

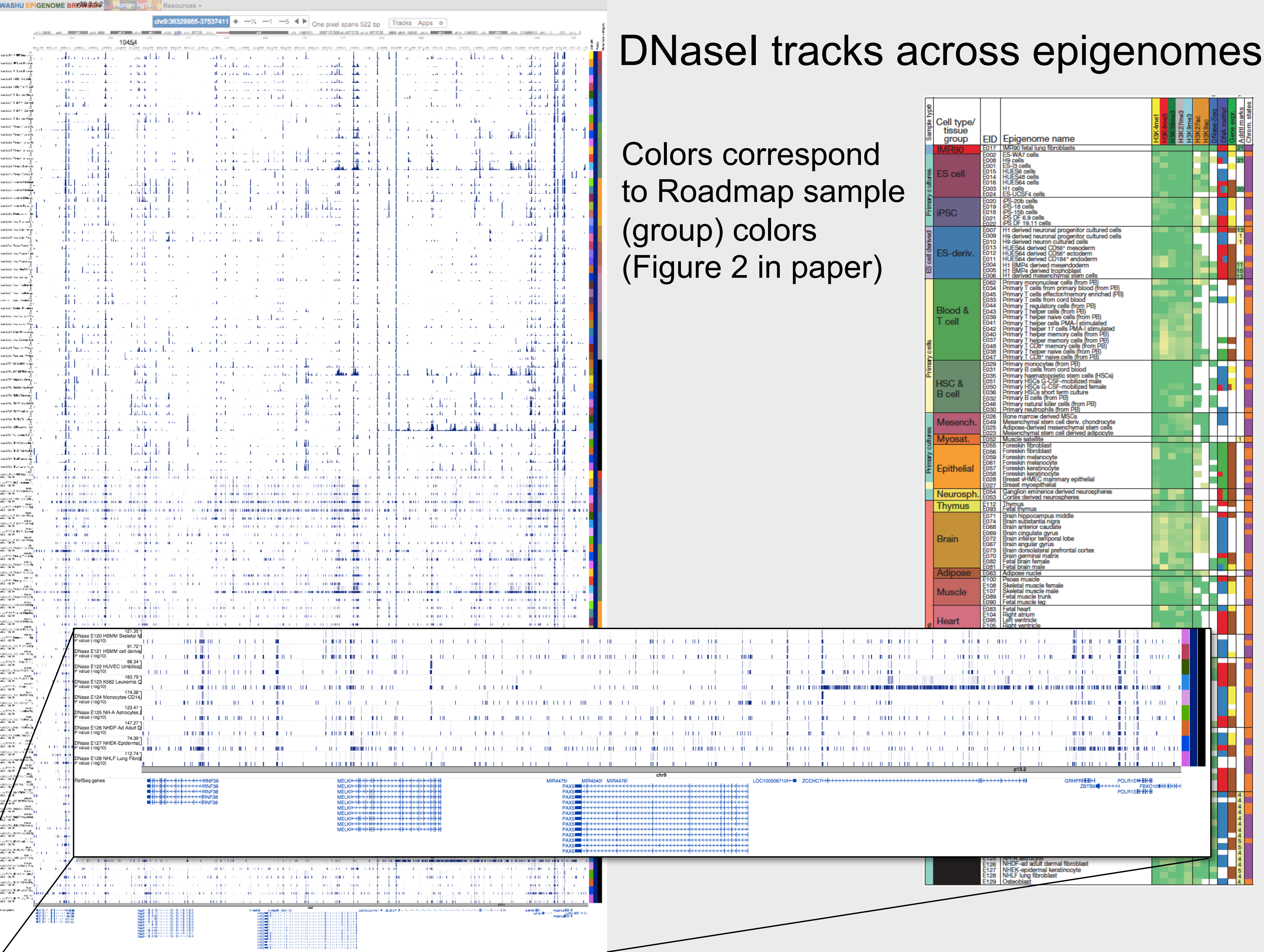
Showing 1 to 127 of 127 entries (filtered from 300 total entries)

Example: all available data for Adipose Nuclei



DNaseI tracks across epigenomes

Colors correspond to Roadmap sample (group) colors (Figure 2 in paper)

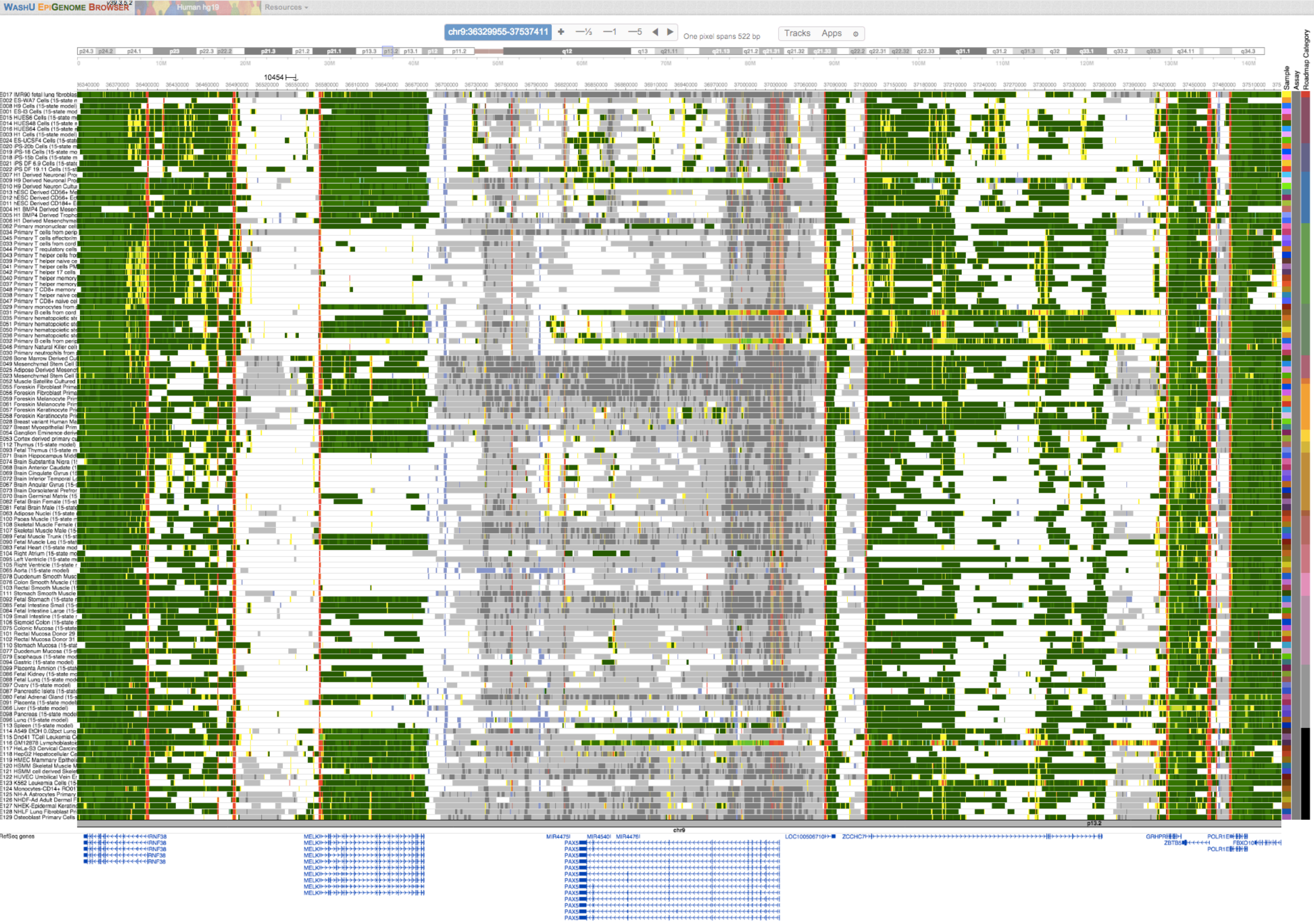


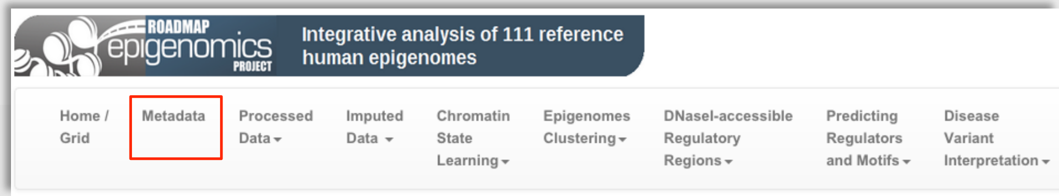
DNase E120 HSMM Skeletal M 121.35 P value (-log10)
DNase E121 HSMM cell derive 91.72 P value (-log10)
DNase E122 HUVEC Umbilical 68.34 P value (-log10)
DNase E123 K562 Leukemia C 183.79 P value (-log10)
DNase E124 Monocytes-CD14 174.38 P value (-log10)
DNase E125 NH-A Astrocytes 123.41 P value (-log10)
DNase E126 NHDF-Ad Adult D 147.27 P value (-log10)
DNase E127 NHEK-Epidermal 74.39 P value (-log10)
DNase E128 NHLF Lung Fibrob 112.74 P value (-log10)

Sample type	Cell type/ tissue group	EID	Epigenome name	Addtl marks
	IMR90	E017	IMR90 fetal lung fibroblasts	21
Primary cultures	ES cell	E002	ES-WA7 cells	
		E008	H9 cells	21
		E001	ES-I3 cells	
		E015	HUES6 cells	
		E014	HUES48 cells	
		E016	HUES64 cells	
		E003	H1 cells	20
		E024	ES-UCSF4 cells	
	iPSC	E020	iPS-20b cells	
		E019	iPS-18 cells	
		E018	iPS-15b cells	
		E021	iPS DF 6.9 cells	
		E022	iPS DF 19.11 cells	
ES cell derived	ES-deriv.	E007	H1 derived neuronal progenitor cultured cells	13
		E009	H9 derived neuronal progenitor cultured cells	1
		E010	H9 derived neuron cultured cells	1
		E013	HUES64 derived CD56⁺ mesoderm	
		E012	HUES64 derived CD56⁺ ectoderm	
		E011	HUES64 derived CD184⁺ endoderm	
		E004	H1 BMP4 derived mesendoderm	11
		E005	H1 BMP4 derived trophoblast	15
		E006	H1 derived mesenchymal stem cells	13
Primary cells	Blood & T cell	E062	Primary mononuclear cells (from PB)	
		E034	Primary T cells from primary blood (from PB)	
		E045	Primary T cells effector/memory enriched (PB)	
		E033	Primary T cells from cord blood	
		E044	Primary T regulatory cells (from PB)	
		E043	Primary T helper cells (from PB)	
		E039	Primary T helper naive cells (from PB)	
		E041	Primary T helper cells PMA-I stimulated	
		E042	Primary T helper 17 cells PMA-I stimulated	
		E040	Primary T helper memory cells (from PB)	
		E037	Primary T helper memory cells (from PB)	
		E048	Primary T CD8⁺ memory cells (from PB)	
		E038	Primary T helper naive cells (from PB)	
		E047	Primary T CD8⁺ naive cells (from PB)	
	HSC & B cell	E029	Primary monocytes (from PB)	
		E031	Primary B cells from cord blood	
		E035	Primary haematopoietic stem cells (HSCs)	
		E051	Primary HSCs G-CSF-mobilized male	
		E050	Primary HSCs G-CSF-mobilized female	
		E036	Primary HSCs short term culture	
		E032	Primary B cells (from PB)	
		E046	Primary natural killer cells (from PB)	
		E030	Primary neutrophils (from PB)	
Primary cultures	Mesench.	E026	Bone marrow derived MSCs	
		E049	Mesenchymal stem cell deriv. chondrocyte	
		E025	Adipose-derived mesenchymal stem cells	
		E023	Mesenchymal stem cell derived adipocyte	
	Myosat.	E052	Muscle satellite	1
	Epithelial	E055	Foreskin fibroblast	
		E056	Foreskin fibroblast	
		E059	Foreskin melanocyte	
		E061	Foreskin melanocyte	
		E057	Foreskin keratinocyte	
		E058	Foreskin keratinocyte	
		E028	Breast vHMEC mammary epithelial	
		E027	Breast myoepithelial	
	Neurosph.	E054	Ganglion eminence derived neurospheres	
		E053	Cortex derived neurospheres	
	Thymus	E112	Thymus	
		E093	Fetal thymus	
	Brain	E071	Brain hippocampus middle	
		E074	Brain substantia nigra	
		E068	Brain anterior caudate	
		E069	Brain cingulate gyrus	
		E072	Brain inferior temporal lobe	
		E067	Brain angular gyrus	
		E073	Brain dorsolateral prefrontal cortex	
		E070	Brain germinal matrix	
		E082	Fetal brain female	
		E081	Fetal brain male	
	Adipose	E063	Adipose nuclei	
	Muscle	E100	Psoas muscle	
		E108	Skeletal muscle female	
		E107	Skeletal muscle male	
		E089	Fetal muscle trunk	
		E090	Fetal muscle leg	
	Heart	E083	Fetal heart	
		E104	Right atrium	
		E095	Left ventricle	
		E105	Right ventricle	
		E125	NH-A astrocyte	4
		E126	NHDF-ad adult dermal fibroblast	4
		E127	NHEK-epidermal keratinocyte	5
		E128	NHLF lung fibroblast	4
		E129	Osteoblast	4

Column headers: H3K4me1, H3K4me3, H3K36me3, H3K27me3, H3K9me3, H3K27ac, H3K9ac, DNase-Seq, DNA methyl, Gene expr., Addtl marks, Chrom. states

15-state chromatin state model (ChromHMM)





Sample naming/IDs, sample grouping, data quality, raw data filenames, metadata

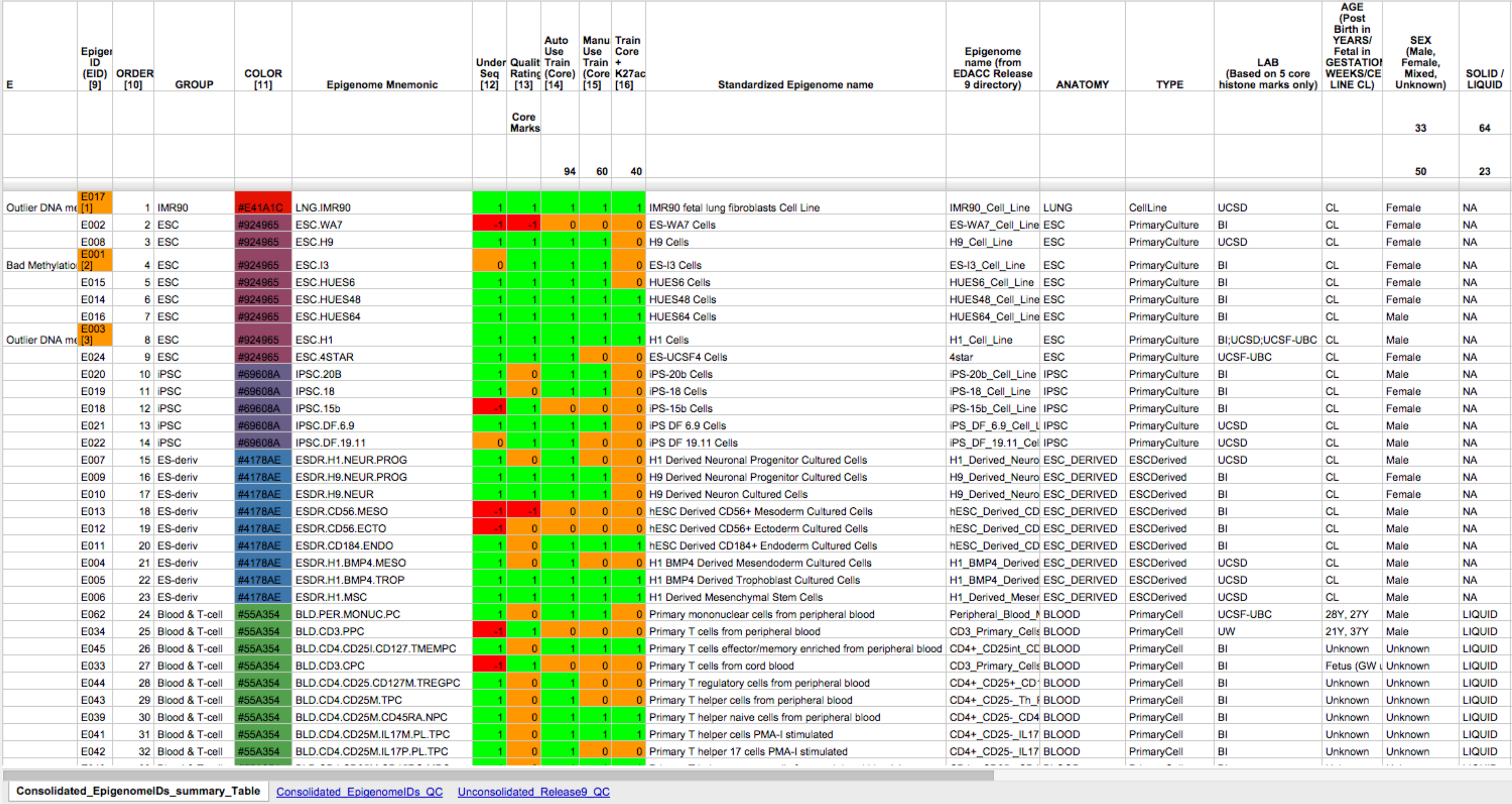
DATA SOURCE

- Google spreadsheet: Metadata and quality control

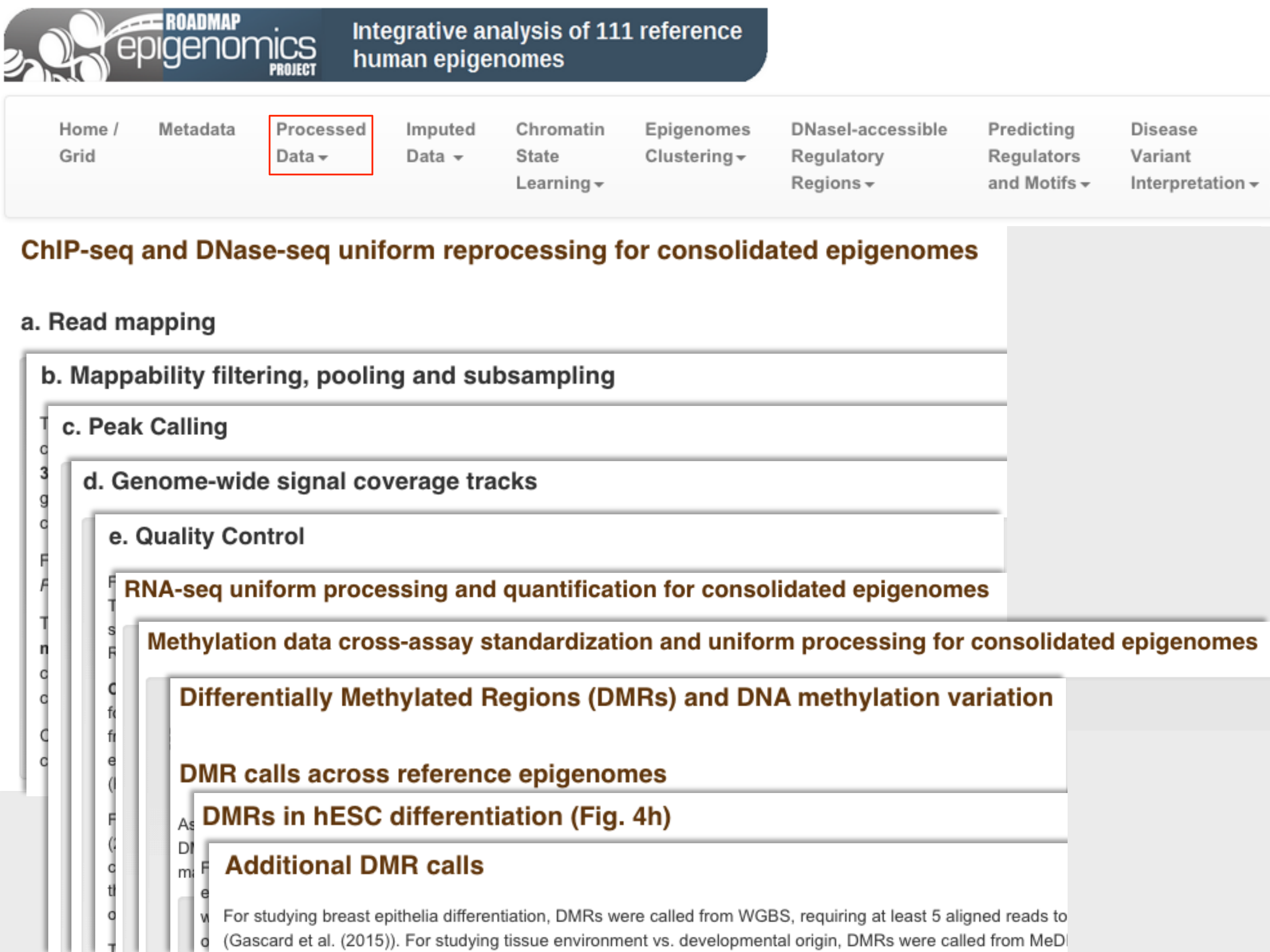
Visualize

The spreadsheet contains 3 sheets (See bottom of sheet)

- **Consolidated_EpigenomeIDs_summary_Table**: Main metadata table for 127 consolidated epigenomes
- **Consolidated_EpigenomeIDs_QC**: QC measures for DNase-seq and Histone ChIP-seq datasets from all 127 consolidated epigenomes
- **Unconsolidated_Release9_QC**: QC measures for DNase-seq and Histone ChIP-seq datasets from all unconsolidated epigenomes
- Comments in the column headers describe each column (scroll over column header if you open the spreadsheet in a new window or see bottom of sheet if you open it using the *Visualize* button)



E	Epiger ID (EID) [9]	ORDER [10]	GROUP	COLOR [11]	Epigenome Mnemonic	Under Seq [12]	Qualit Rating [13]	Auto Use Train (Core) [14]	Manu Use Train (Core [15]	Train Core + K27ac [16]	Standardized Epigenome name	Epigenome name (from EDACC Release 9 directory)	ANATOMY	TYPE	LAB (Based on 5 core histone marks only)	AGE (Post Birth in YEARS/ Fetal in GESTATIOI WEEKS/CE LINE CL)	SEX (Male, Female, Mixed, Unknown)	SOLID / LIQUID
							Core Marks											
																	33	64
								94	60	40							50	23
Outlier DNA me	E017 [1]	1	IMR90	#E41A1C	LNG.IMR90	1	1	1	1	1	IMR90 fetal lung fibroblasts Cell Line	IMR90_Cell_Line	LUNG	CellLine	UCSD	CL	Female	NA
	E002	2	ESC	#924965	ESC.WA7	-1	-1	0	0	0	ES-WA7 Cells	ES-WA7_Cell_Line	ESC	PrimaryCulture	BI	CL	Female	NA
	E008	3	ESC	#924965	ESC.H9	1	1	1	1	0	H9 Cells	H9_Cell_Line	ESC	PrimaryCulture	UCSD	CL	Female	NA
Bad Methylatio	E001 [2]	4	ESC	#924965	ESC.I3	0	1	1	1	0	ES-I3 Cells	ES-I3_Cell_Line	ESC	PrimaryCulture	BI	CL	Female	NA
	E015	5	ESC	#924965	ESC.HUES6	1	1	1	1	0	HUES6 Cells	HUES6_Cell_Line	ESC	PrimaryCulture	BI	CL	Female	NA
	E014	6	ESC	#924965	ESC.HUES48	1	1	1	1	1	HUES48 Cells	HUES48_Cell_Line	ESC	PrimaryCulture	BI	CL	Female	NA
	E016	7	ESC	#924965	ESC.HUES64	1	1	1	1	1	HUES64 Cells	HUES64_Cell_Line	ESC	PrimaryCulture	BI	CL	Male	NA
Outlier DNA me	E003 [3]	8	ESC	#924965	ESC.H1	1	1	1	1	1	H1 Cells	H1_Cell_Line	ESC	PrimaryCulture	BI;UCSD;UCSF-UBC	CL	Male	NA
	E024	9	ESC	#924965	ESC.4STAR	1	1	1	0	0	ES-UCSF4 Cells	4star	ESC	PrimaryCulture	UCSF-UBC	CL	Female	NA
	E020	10	iPSC	#69608A	IPSC.20B	1	0	1	1	0	iPS-20b Cells	iPS-20b_Cell_Line	IPSC	PrimaryCulture	BI	CL	Male	NA
	E019	11	iPSC	#69608A	IPSC.18	1	0	1	1	0	iPS-18 Cells	iPS-18_Cell_Line	IPSC	PrimaryCulture	BI	CL	Female	NA
	E018	12	iPSC	#69608A	IPSC.15b	-1	1	0	0	0	iPS-15b Cells	iPS-15b_Cell_Line	IPSC	PrimaryCulture	BI	CL	Female	NA
	E021	13	iPSC	#69608A	IPSC.DF.6.9	1	1	1	1	0	iPS DF 6.9 Cells	iPS_DF_6.9_Cell_L	IPSC	PrimaryCulture	UCSD	CL	Male	NA
	E022	14	iPSC	#69608A	IPSC.DF.19.11	0	1	1	0	0	iPS DF 19.11 Cells	iPS_DF_19.11_Cel	IPSC	PrimaryCulture	UCSD	CL	Male	NA
	E007	15	ES-deriv	#4178AE	ESDR.H1.NEUR.PROG	1	0	1	0	0	H1 Derived Neuronal Progenitor Cultured Cells	H1_Derived_Neuro	ESC_DERIVED	ESCDerived	UCSD	CL	Male	NA
	E009	16	ES-deriv	#4178AE	ESDR.H9.NEUR.PROG	1	1	1	1	0	H9 Derived Neuronal Progenitor Cultured Cells	H9_Derived_Neuro	ESC_DERIVED	ESCDerived	BI	CL	Female	NA
	E010	17	ES-deriv	#4178AE	ESDR.H9.NEUR	1	1	1	1	0	H9 Derived Neuron Cultured Cells	H9_Derived_Neuro	ESC_DERIVED	ESCDerived	BI	CL	Female	NA
	E013	18	ES-deriv	#4178AE	ESDR.CD56.MESO	-1	-1	0	0	0	hESC Derived CD56+ Mesoderm Cultured Cells	hESC_Derived_CD	ESC_DERIVED	ESCDerived	BI	CL	Male	NA
	E012	19	ES-deriv	#4178AE	ESDR.CD56.ECTO	-1	0	0	0	0	hESC Derived CD56+ Ectoderm Cultured Cells	hESC_Derived_CD	ESC_DERIVED	ESCDerived	BI	CL	Male	NA
	E011	20	ES-deriv	#4178AE	ESDR.CD184.ENDO	1	0	1	1	1	hESC Derived CD184+ Endoderm Cultured Cells	hESC_Derived_CD	ESC_DERIVED	ESCDerived	BI	CL	Male	NA
	E004	21	ES-deriv	#4178AE	ESDR.H1.BMP4.MESO	1	0	1	0	0	H1 BMP4 Derived Mesendoderm Cultured Cells	H1_BMP4_Derived	ESC_DERIVED	ESCDerived	UCSD	CL	Male	NA
	E005	22	ES-deriv	#4178AE	ESDR.H1.BMP4.TROP	1	1	1	1	1	H1 BMP4 Derived Trophoblast Cultured Cells	H1_BMP4_Derived	ESC_DERIVED	ESCDerived	UCSD	CL	Male	NA
	E006	23	ES-deriv	#4178AE	ESDR.H1.MSC	1	1	1	1	1	H1 Derived Mesenchymal Stem Cells	H1_Derived_Meser	ESC_DERIVED	ESCDerived	UCSD	CL	Male	NA
	E062	24	Blood & T-cell	#55A354	BLD.PER.MONUC.PC	1	0	1	1	0	Primary mononuclear cells from peripheral blood	Peripheral_Blood_I	BLOOD	PrimaryCell	UCSF-UBC	28Y, 27Y	Male	LIQUID
	E034	25	Blood & T-cell	#55A354	BLD.CD3.PPC	-1	1	0	0	0	Primary T cells from peripheral blood	CD3_Primary_Cells	BLOOD	PrimaryCell	UW	21Y, 37Y	Male	LIQUID
	E045	26	Blood & T-cell	#55A354	BLD.CD4.CD25I.CD127.TMEMPC	1	0	1	1	1	Primary T cells effector/memory enriched from peripheral blood	CD4+_CD25int_CD	BLOOD	PrimaryCell	BI	Unknown	Unknown	LIQUID
	E033	27	Blood & T-cell	#55A354	BLD.CD3.CPC	-1	1	0	0	0	Primary T cells from cord blood	CD3_Primary_Cells	BLOOD	PrimaryCell	BI	Fetus (GW u	Unknown	LIQUID
	E044	28	Blood & T-cell	#55A354	BLD.CD4.CD25.CD127M.TREGPC	1	0	1	0	0	Primary T regulatory cells from peripheral blood	CD4+_CD25+_CD1	BLOOD	PrimaryCell	BI	Unknown	Unknown	LIQUID
	E043	29	Blood & T-cell	#55A354	BLD.CD4.CD25M.TPC	1	0	1	0	0	Primary T helper cells from peripheral blood	CD4+_CD25-_Th_I	BLOOD	PrimaryCell	BI	Unknown	Unknown	LIQUID
	E039	30	Blood & T-cell	#55A354	BLD.CD4.CD25M.CD45RA.NPC	1	0	1	1	1	Primary T helper naive cells from peripheral blood	CD4+_CD25-_CD4	BLOOD	PrimaryCell	BI	Unknown	Unknown	LIQUID
	E041	31	Blood & T-cell	#55A354	BLD.CD4.CD25M.IL17M.PL.TPC	1	0	1	1	1	Primary T helper cells PMA-I stimulated	CD4+_CD25-_IL17	BLOOD	PrimaryCell	BI	Unknown	Unknown	LIQUID
	E042	32	Blood & T-cell	#55A354	BLD.CD4.CD25M.IL17P.PL.TPC	1	0	1	0	0	Primary T helper 17 cells PMA-I stimulated	CD4+_CD25-_IL17	BLOOD	PrimaryCell	BI	Unknown	Unknown	LIQUID



ChIP-seq and DNase-seq uniform reprocessing for consolidated epigenomes

a. Read mapping

b. Mappability filtering, pooling and subsampling

c. Peak Calling

d. Genome-wide signal coverage tracks

e. Quality Control

RNA-seq uniform processing and quantification for consolidated epigenomes

Methylation data cross-assay standardization and uniform processing for consolidated epigenomes

Differentially Methylated Regions (DMRs) and DNA methylation variation

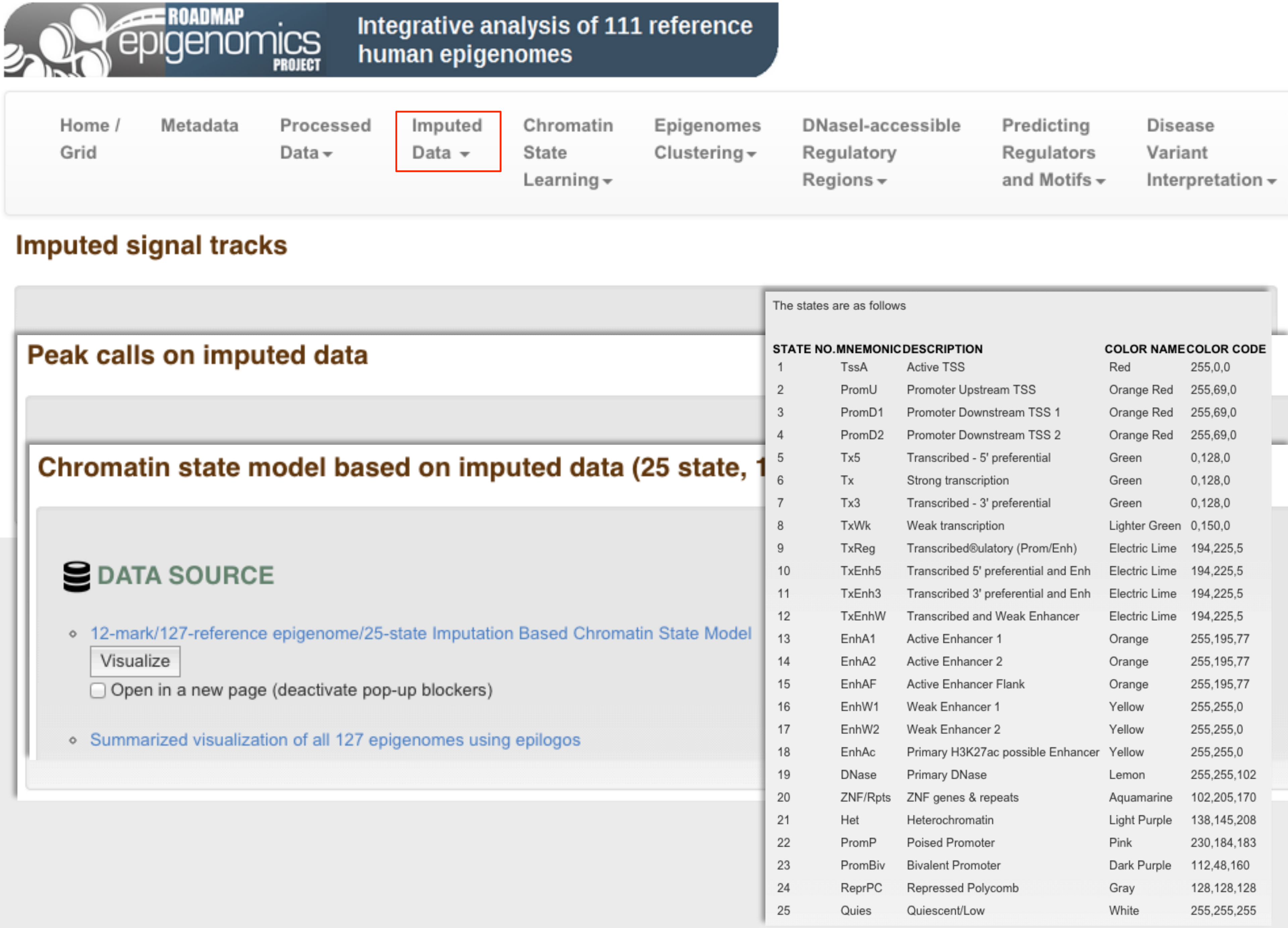
DMR calls across reference epigenomes

DMRs in hESC differentiation (Fig. 4h)

Additional DMR calls

For studying breast epithelia differentiation, DMRs were called from WGBS, requiring at least 5 aligned reads to

(Gascard et al. (2015)). For studying tissue environment vs. developmental origin, DMRs were called from MeD



Integrative analysis of 111 reference human epigenomes

Imputed signal tracks

Peak calls on imputed data

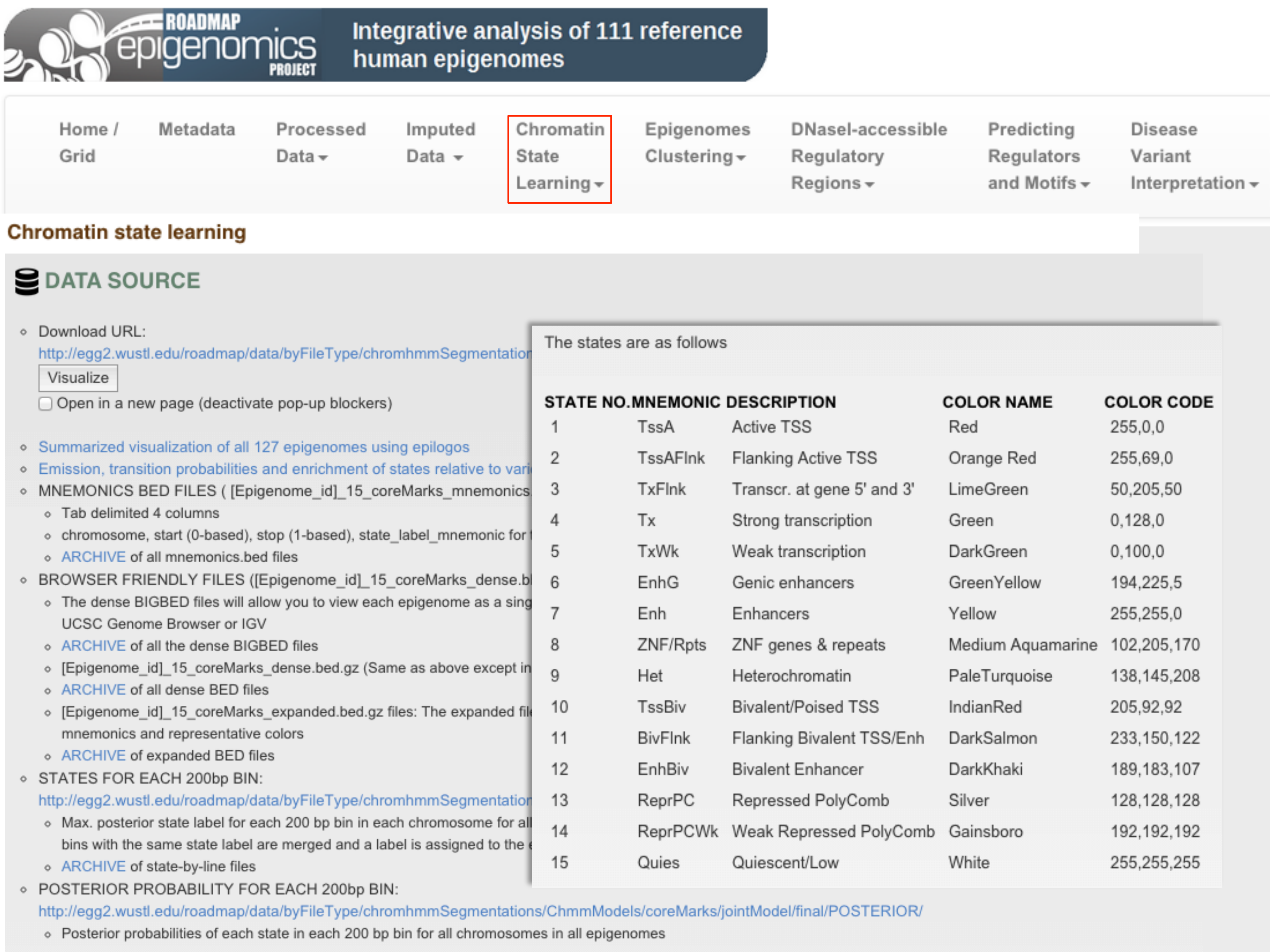
Chromatin state model based on imputed data (25 state, 1

DATA SOURCE

- 12-mark/127-reference epigenome/25-state Imputation Based Chromatin State Model
  ☐ Open in a new page (deactivate pop-up blockers)
- Summarized visualization of all 127 epigenomes using epilogos

The states are as follows

STATE NO.	MNEMONIC	DESCRIPTION	COLOR NAME	COLOR CODE
1	TssA	Active TSS	Red	255,0,0
2	PromU	Promoter Upstream TSS	Orange Red	255,69,0
3	PromD1	Promoter Downstream TSS 1	Orange Red	255,69,0
4	PromD2	Promoter Downstream TSS 2	Orange Red	255,69,0
5	Tx5	Transcribed - 5' preferential	Green	0,128,0
6	Tx	Strong transcription	Green	0,128,0
7	Tx3	Transcribed - 3' preferential	Green	0,128,0
8	TxWk	Weak transcription	Lighter Green	0,150,0
9	TxReg	Transcribed®ulatory (Prom/Enh)	Electric Lime	194,225,5
10	TxEnh5	Transcribed 5' preferential and Enh	Electric Lime	194,225,5
11	TxEnh3	Transcribed 3' preferential and Enh	Electric Lime	194,225,5
12	TxEnhW	Transcribed and Weak Enhancer	Electric Lime	194,225,5
13	EnhA1	Active Enhancer 1	Orange	255,195,77
14	EnhA2	Active Enhancer 2	Orange	255,195,77
15	EnhAF	Active Enhancer Flank	Orange	255,195,77
16	EnhW1	Weak Enhancer 1	Yellow	255,255,0
17	EnhW2	Weak Enhancer 2	Yellow	255,255,0
18	EnhAc	Primary H3K27ac possible Enhancer	Yellow	255,255,0
19	DNase	Primary DNase	Lemon	255,255,102
20	ZNF/Rpts	ZNF genes & repeats	Aquamarine	102,205,170
21	Het	Heterochromatin	Light Purple	138,145,208
22	PromP	Poised Promoter	Pink	230,184,183
23	PromBiv	Bivalent Promoter	Dark Purple	112,48,160
24	ReprPC	Repressed Polycomb	Gray	128,128,128
25	Quies	Quiescent/Low	White	255,255,255



Integrative analysis of 111 reference human epigenomes

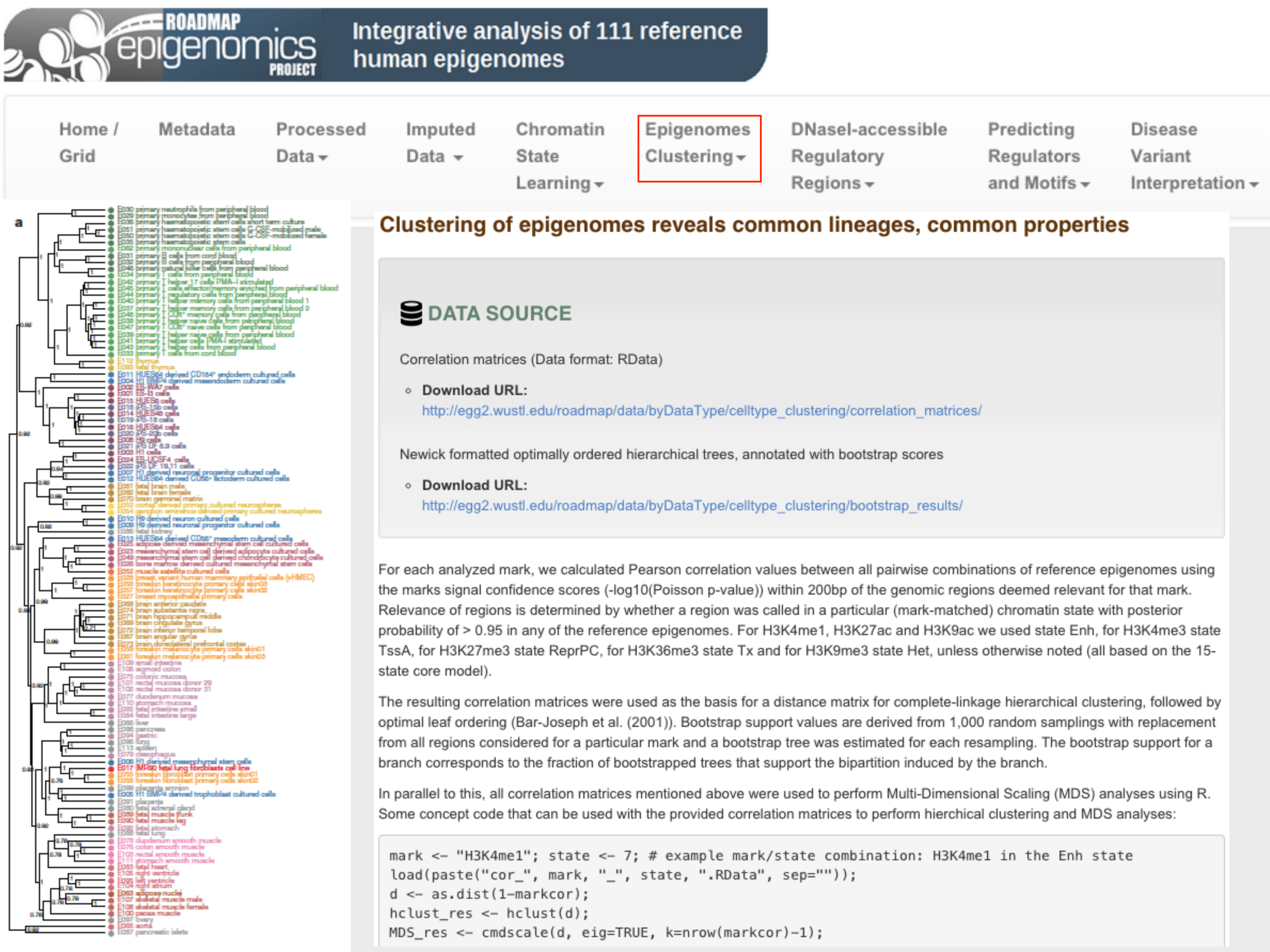
Chromatin state learning

DATA SOURCE

- Download URL:
  http://egg2.wustl.edu/roadmap/data/byFileType/chromhmmSegmentation
  ☐ Open in a new page (deactivate pop-up blockers)
- Summarized visualization of all 127 epigenomes using epilogos
- Emission, transition probabilities and enrichment of states relative to vari
- MNEMONICS BED FILES ([Epigenome_id]_15_coreMarks_mnemonics
  - Tab delimited 4 columns
  - chromosome, start (0-based), stop (1-based), state_label_mnemonic for
  - ARCHIVE of all mnemonics.bed files
- BROWSER FRIENDLY FILES ([Epigenome_id]_15_coreMarks_dense.bi
  - The dense BIGBED files will allow you to view each epigenome as a sing UCSC Genome Browser or IGV
  - ARCHIVE of all the dense BIGBED files
  - [Epigenome_id]_15_coreMarks_dense.bed.gz (Same as above except in
  - ARCHIVE of all dense BED files
  - [Epigenome_id]_15_coreMarks_expanded.bed.gz files: The expanded fil mnemonics and representative colors
  - ARCHIVE of expanded BED files
- STATES FOR EACH 200bp BIN:
  http://egg2.wustl.edu/roadmap/data/byFileType/chromhmmSegmentatio
  - Max. posterior state label for each 200 bp bin in each chromosome for all bins with the same state label are merged and a label is assigned to the e
  - ARCHIVE of state-by-line files
- POSTERIOR PROBABILITY FOR EACH 200bp BIN:
  http://egg2.wustl.edu/roadmap/data/byFileType/chromhmmSegmentations/ChmmModels/coreMarks/jointModel/final/POSTERIOR/
  - Posterior probabilities of each state in each 200 bp bin for all chromosomes in all epigenomes

The states are as follows

STATE NO.	MNEMONIC	DESCRIPTION	COLOR NAME	COLOR CODE
1	TssA	Active TSS	Red	255,0,0
2	TssAFlnk	Flanking Active TSS	Orange Red	255,69,0
3	TxFlnk	Transcr. at gene 5' and 3'	LimeGreen	50,205,50
4	Tx	Strong transcription	Green	0,128,0
5	TxWk	Weak transcription	DarkGreen	0,100,0
6	EnhG	Genic enhancers	GreenYellow	194,225,5
7	Enh	Enhancers	Yellow	255,255,0
8	ZNF/Rpts	ZNF genes & repeats	Medium Aquamarine	102,205,170
9	Het	Heterochromatin	PaleTurquoise	138,145,208
10	TssBiv	Bivalent/Poised TSS	IndianRed	205,92,92
11	BivFlnk	Flanking Bivalent TSS/Enh	DarkSalmon	233,150,122
12	EnhBiv	Bivalent Enhancer	DarkKhaki	189,183,107
13	ReprPC	Repressed PolyComb	Silver	128,128,128
14	ReprPCWk	Weak Repressed PolyComb	Gainsboro	192,192,192
15	Quies	Quiescent/Low	White	255,255,255



Integrative analysis of 111 reference human epigenomes

Clustering of epigenomes reveals common lineages, common properties

DATA SOURCE

Correlation matrices (Data format: RData)

- **Download URL:** http://egg2.wustl.edu/roadmap/data/byDataType/celltype_clustering/correlation_matrices/

Newick formatted optimally ordered hierarchical trees, annotated with bootstrap scores

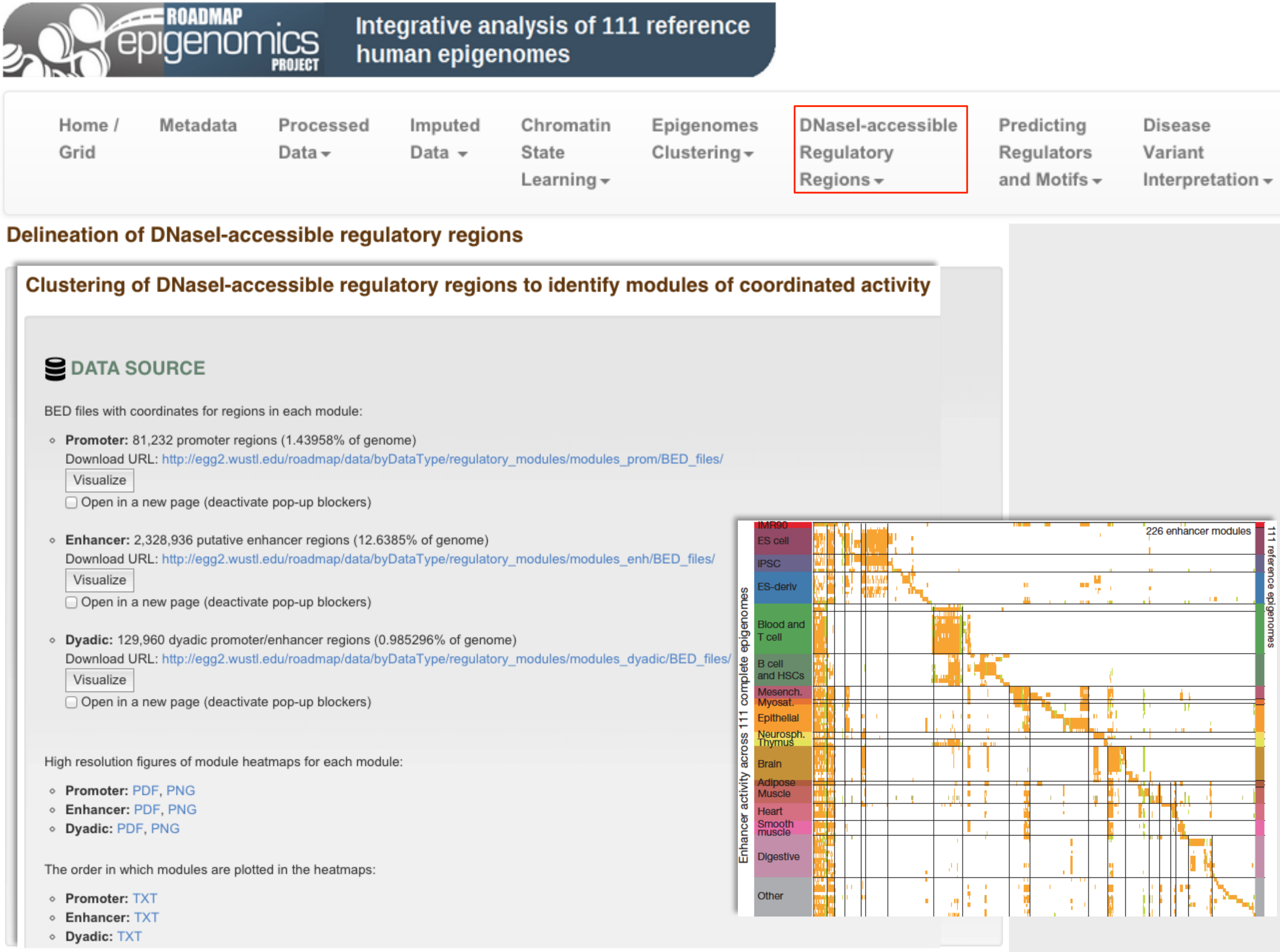
- **Download URL:** http://egg2.wustl.edu/roadmap/data/byDataType/celltype_clustering/bootstrap_results/

For each analyzed mark, we calculated Pearson correlation values between all pairwise combinations of reference epigenomes using the marks signal confidence scores (-log10(Poisson p-value)) within 200bp of the genomic regions deemed relevant for that mark. Relevance of regions is determined by whether a region was called in a particular (mark-matched) chromatin state with posterior probability of > 0.95 in any of the reference epigenomes. For H3K4me1, H3K27ac and H3K9ac we used state Enh, for H3K4me3 state TssA, for H3K27me3 state ReprPC, for H3K36me3 state Tx and for H3K9me3 state Het, unless otherwise noted (all based on the 15-state core model).

The resulting correlation matrices were used as the basis for a distance matrix for complete-linkage hierarchical clustering, followed by optimal leaf ordering (Bar-Joseph et al. (2001)). Bootstrap support values are derived from 1,000 random samplings with replacement from all regions considered for a particular mark and a bootstrap tree was estimated for each resampling. The bootstrap support for a branch corresponds to the fraction of bootstrapped trees that support the bipartition induced by the branch.

In parallel to this, all correlation matrices mentioned above were used to perform Multi-Dimensional Scaling (MDS) analyses using R. Some concept code that can be used with the provided correlation matrices to perform hierchical clustering and MDS analyses:

```
mark <- "H3K4me1"; state <- 7; # example mark/state combination: H3K4me1 in the Enh state
load(paste("cor_", mark, "_", state, ".RData", sep=""));
d <- as.dist(1-markcor);
hclust_res <- hclust(d);
MDS_res <- cmdscale(d, eig=TRUE, k=nrow(markcor)-1);
```



Integrative analysis of 111 reference human epigenomes

Delineation of DNaseI-accessible regulatory regions

Clustering of DNaseI-accessible regulatory regions to identify modules of coordinated activity

DATA SOURCE

BED files with coordinates for regions in each module:

- **Promoter:** 81,232 promoter regions (1.43958% of genome)
  Download URL: http://egg2.wustl.edu/roadmap/data/byDataType/regulatory_modules/modules_prom/BED_files/
  Visualize
  ☐ Open in a new page (deactivate pop-up blockers)
- **Enhancer:** 2,328,936 putative enhancer regions (12.6385% of genome)
  Download URL: http://egg2.wustl.edu/roadmap/data/byDataType/regulatory_modules/modules_enh/BED_files/
  Visualize
  ☐ Open in a new page (deactivate pop-up blockers)
- **Dyadic:** 129,960 dyadic promoter/enhancer regions (0.985296% of genome)
  Download URL: http://egg2.wustl.edu/roadmap/data/byDataType/regulatory_modules/modules_dyadic/BED_files/
  Visualize
  ☐ Open in a new page (deactivate pop-up blockers)

High resolution figures of module heatmaps for each module:

- **Promoter:** PDF, PNG
- **Enhancer:** PDF, PNG
- **Dyadic:** PDF, PNG

The order in which modules are plotted in the heatmaps:

- **Promoter:** TXT
- **Enhancer:** TXT
- **Dyadic:** TXT

ROADMAP epigenomics PROJECT

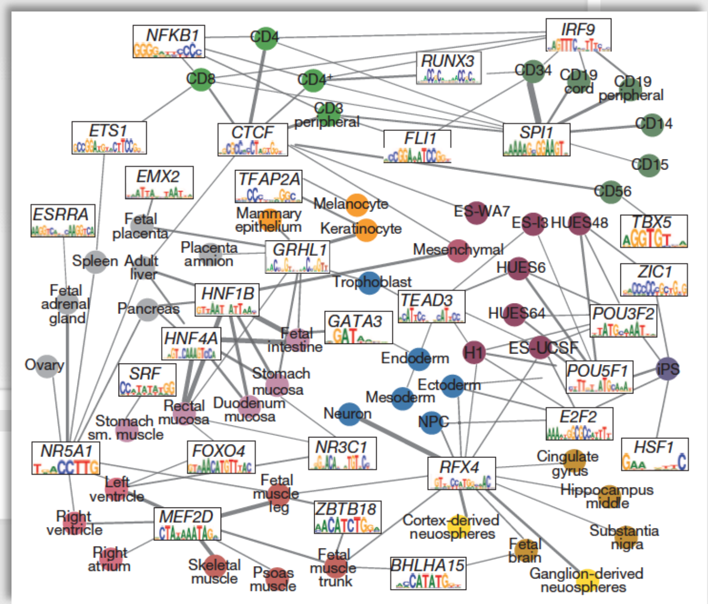
Integrative analysis of 111 reference human epigenomes

Home / Grid | Metadata | Processed Data ▾ | Imputed Data ▾ | Chromatin State Learning ▾ | Epigenomes Clustering ▾ | DNaseI-accessible Regulatory Regions ▾ | Predicting Regulators and Motifs ▾ | Disease Variant Interpretation ▾

Predicting motifs and active regulators in each cell-type/tissue/lineage

DATA SOURCE

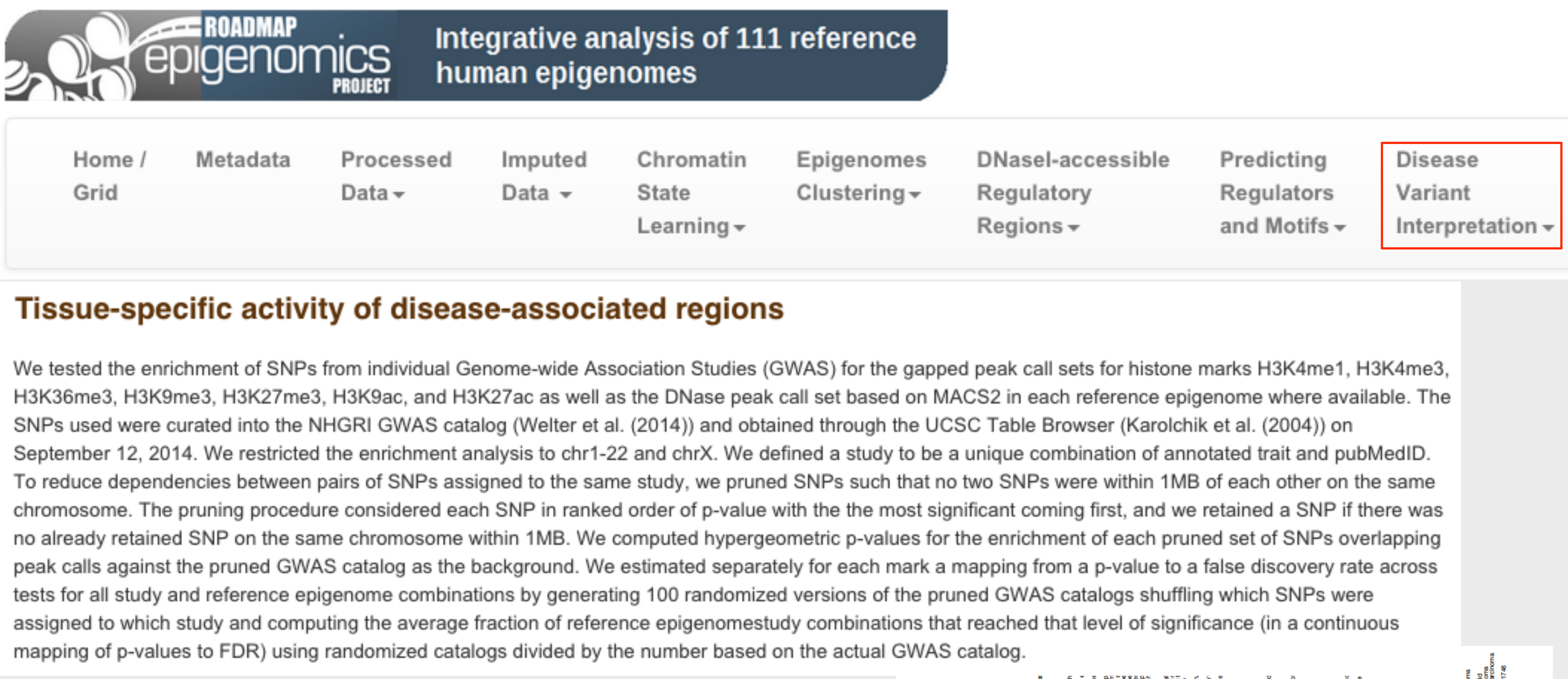
- Primary motif resource:
  http://compbio.mit.edu/encode-motifs/
- Data underlying Figures Extended Data 8a, S13a (clustered), S13b,c (unclustered), S13d,e (unique):
  http://egg2.wustl.edu/roadmap/data/byDataType/motifanalysis/pouyak/
- Motif enrichments (enhancer cluster centric):
  http://egg2.wustl.edu/roadmap/data/byDataType/motifanalysis/pouyak/viewByCluster/bycluster.html



DNA Motif Positional Bias in Digital Genomic Footprinting Sites

DATA SOURCE

- Primary motif resource:
  http://compbio.mit.edu/encode-motifs/
- DNase/DGF Footprint calls:
  http://egg2.wustl.edu/roadmap/data/byDataType/dgfootprints/
  Format: 5 column BED files. 4th column is footprint ID. 5th column is FOS score. See below.
- Data underlying Figures 8, Extended Data 9b, 9c:
  http://egg2.wustl.edu/roadmap/data/byDataType/motifanalysis/zhizhou/

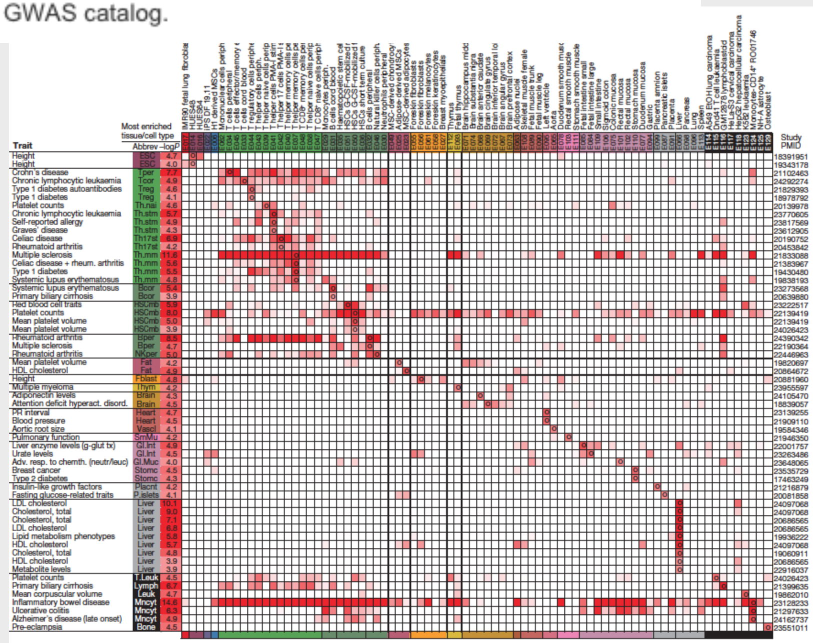


Integrative analysis of 111 reference human epigenomes

Tissue-specific activity of disease-associated regions

We tested the enrichment of SNPs from individual Genome-wide Association Studies (GWAS) for the gapped peak call sets for histone marks H3K4me1, H3K4me3, H3K36me3, H3K9me3, H3K27me3, H3K9ac, and H3K27ac as well as the DNase peak call set based on MACS2 in each reference epigenome where available. The SNPs used were curated into the NHGRI GWAS catalog (Welter et al. (2014)) and obtained through the UCSC Table Browser (Karolchik et al. (2004)) on September 12, 2014. We restricted the enrichment analysis to chr1-22 and chrX. We defined a study to be a unique combination of annotated trait and pubMedID. To reduce dependencies between pairs of SNPs assigned to the same study, we pruned SNPs such that no two SNPs were within 1MB of each other on the same chromosome. The pruning procedure considered each SNP in ranked order of p-value with the the most significant coming first, and we retained a SNP if there was no already retained SNP on the same chromosome within 1MB. We computed hypergeometric p-values for the enrichment of each pruned set of SNPs overlapping peak calls against the pruned GWAS catalog as the background. We estimated separately for each mark a mapping from a p-value to a false discovery rate across tests for all study and reference epigenome combinations by generating 100 randomized versions of the pruned GWAS catalogs shuffling which SNPs were assigned to which study and computing the average fraction of reference epigenomestudy combinations that reached that level of significance (in a continuous mapping of p-values to FDR) using randomized catalogs divided by the number based on the actual GWAS catalog.

Further data available in Table S6 online



Integrative analysis of 111 reference human epigenomes

- Home / Grid
- Metadata
- Processed Data
- Imputed Data
- Chromatin State Learning
- Epigenomes Clustering
- DNaseI-accessible Regulatory Regions
- Predicting Regulators and Motifs
- Disease Variant Interpretation

DATA SOURCE

- Download URL:
  http://egg2.wustl.edu/roadmap/data/byFileType/chromhmmSegmentations/ChmmModels/coreMarks/jointModel/final
  Visualize
  ☐ Open in a new page (deactivate pop-up blockers)
- Summarized visualization of all 127 epigenomes using epilogos

epilogos